Indian Liver Patient Records

Wilson Moreno

1/8/2020

Introduction

The present work wishes to predict if a patient in India is going to suffer from liver disease, by means of some indicators, for example: Age, Gender, Alkaline_Phosphography, etc. The result of a categorical variable. Different types of approaches are used such as: using all variables, through a correlation matrix and by specific variables.

Methods/Analysis

Step 0: require package

```
if (!require(package)) install.packages('psych', repos = "http://cran.us.r-project.org")
if (!require(package)) install.packages('knitr', repos = "http://cran.us.r-project.org")
if (!require(package)) install.packages('ggplot2', repos = "http://cran.us.r-project.org")
library(knitr)
library(ggplot2)
library(psych)
library(caret)
```

Step 1: Load the data base

```
database <- read.csv("Data/indian_liver_patient.csv")
```

Step 2: Exploratory Data Analysis

Summay Statistics

The following table shows the descriptive statistics for all the variables in the database.

```
round(data.frame( describeBy(database, digits= 2)),1)
```

```
## Warning in describeBy(database, digits = 2): no grouping variable requeste
d
```

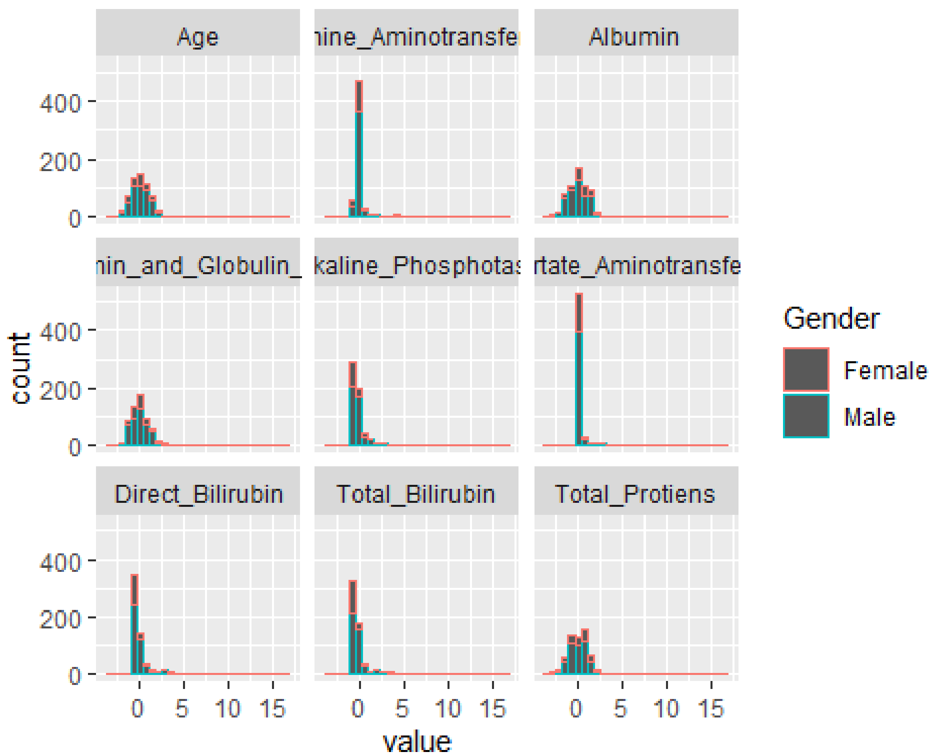
```
##                             vars   n  mean    sd median trimmed  mad  min
  max
## Age                            1 583  44.7  16.2   45.0    44.8 17.8  4.0
 90.0
## Gender*                        2 583   1.8   0.4    2.0     1.8  0.0  1.0
  2.0
## Total_Bilirubin                3 583   3.3   6.2    1.0     1.7  0.4  0.4
 75.0
## Direct_Bilirubin               4 583   1.5   2.8    0.3     0.7  0.3  0.1
 19.7
## Alkaline_Phosphotase           5 583 290.6 242.9  208.0   238.4 74.1 63.0 2
110.0
## Alamine_Aminotransferase       6 583  80.7 182.6   35.0    43.9 22.2 10.0 2
000.0
## Aspartate_Aminotransferase     7 583 109.9 288.9   42.0    56.8 31.1 10.0 4
929.0
## Total_Protiens                 8 583   6.5   1.1    6.6     6.5  1.0  2.7
  9.6
## Albumin                        9 583   3.1   0.8    3.1     3.1  0.9  0.9
  5.5
## Albumin_and_Globulin_Ratio    10 579   0.9   0.3    0.9     0.9  0.3  0.3
  2.8
## Dataset                       11 583   1.3   0.5    1.0     1.2  0.0  1.0
  2.0
##                             range skew kurtosis   se
## Age                          86.0  0.0     -0.6  0.7
## Gender*                       1.0 -1.2     -0.6  0.0
## Total_Bilirubin              74.6  4.9     36.7  0.3
## Direct_Bilirubin             19.6  3.2     11.2  0.1
## Alkaline_Phosphotase       2047.0  3.7     17.5 10.1
## Alamine_Aminotransferase   1990.0  6.5     50.0  7.6
## Aspartate_Aminotransferase 4919.0 10.5    149.1 12.0
## Total_Protiens                6.9 -0.3      0.2  0.0
## Albumin                       4.6  0.0     -0.4  0.0
## Albumin_and_Globulin_Ratio    2.5  1.0      3.2  0.0
## Dataset                       1.0  0.9     -1.1  0.0
```

```
# tmp <- describeBy(database,
#             group = database$Gender,
#             digits= 1)
```

Visualization

As can be seen in the visual analysis of the data, they were divided into 2 groups by gender, in all the graphs the rates are much higher in women than in men, and tend to follow the same distribution. Variables: Age, Albumin, Albumin_and_Globulin_Ratio and Total_Protiens, are suspected to follow a normal distribution.

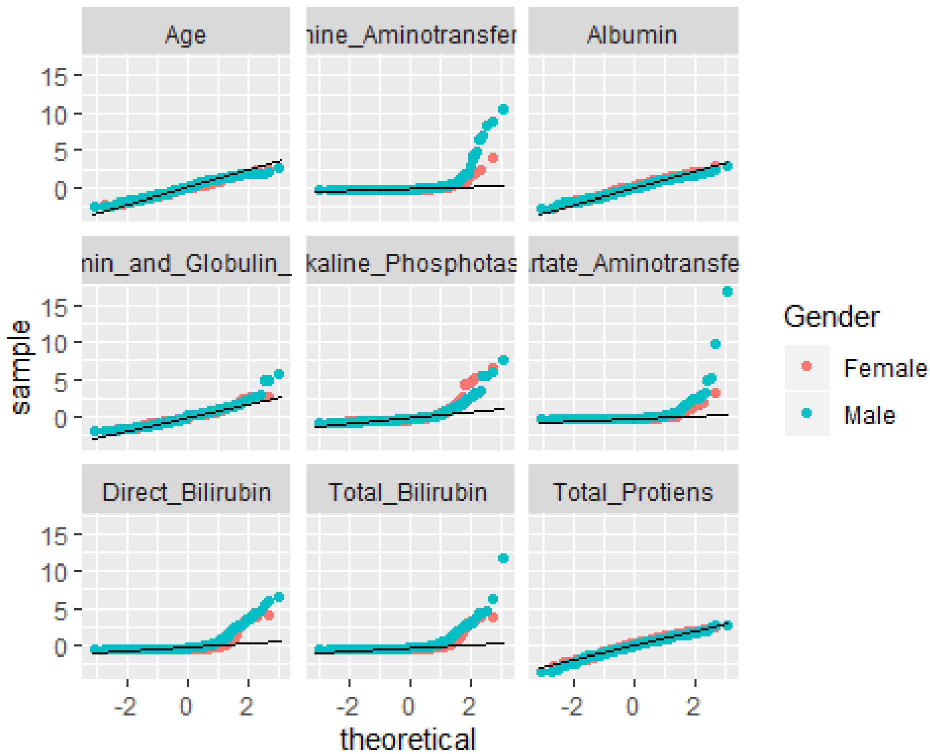
```
scale_database <- database %>% select(-Gender,-Dataset) %>% scale() %>% as.da
ta.frame() %>%
  cbind(Gender = database$Gender)

database.gathered <- scale_database  %>% as.data.frame() %>%
  gather(key = "variable", value = "value", - Gender)

ggplot(data = database.gathered , mapping = aes(x = value, color = Gender)) +
  geom_histogram() +
  facet_wrap(facets =  vars(variable ))
```



This graph certifies the suspicion that the aforementioned variables follow a normal distribution.

```
ggplot(data = database.gathered , mapping = aes(sample = value, color = Gende
r)) +
```

```
  stat_qq() +  stat_qq_line(color = "black") +
  facet_wrap(facets =  vars(variable ))
```



Step 3: Split the database in training and testing

Split the database in training and testing

```
set.seed(755)
test_index <- createDataPartition(y = database$Dataset, times = 1,
                                  p = 0.2, list = FALSE)

train_set <- database[-test_index,]
test_set <- database[test_index,]


RMSE <- function(true_ratings, predicted_ratings){
  sqrt(mean((true_ratings - predicted_ratings)^2))
}
```

How will the logistic regression algorithm be used **glm ()** you have to modify the response variable **Data set** to values of 0 if the person does not suffer the disease and 1 if a person suffers the disease and convert the gender variable into a factor.

```
# Modify the response variable.
train_set$Dataset[train_set$Dataset == 1] = 1
```

```r
train_set$Dataset[train_set$Dataset == 2] = 0
test_set$Dataset[test_set$Dataset == 1] = 1
test_set$Dataset[test_set$Dataset == 2] = 0

# convert the variable into factor
train_set$Gender <- as.factor(train_set$Gender)
test_set$Gender <- as.factor(test_set$Gender)
```

Step 4: Choose the Model and train the model with the training base

For the following approach some models are used, which are a function of the predictive variables to choose.

Model 1: Using all the variables

```r
mol_1 <- glm(Dataset ~. , data = train_set, family = binomial() )
```

Model 2: Variables + correlation

```r
round(cor(database[,-2]),2)

##                              Age Total_Bilirubin Direct_Bilirubin
## Age                         1.00            0.01             0.01
## Total_Bilirubin             0.01            1.00             0.87
## Direct_Bilirubin            0.01            0.87             1.00
## Alkaline_Phosphotase        0.08            0.21             0.23
## Alamine_Aminotransferase   -0.09            0.21             0.23
## Aspartate_Aminotransferase -0.02            0.24             0.26
## Total_Protiens             -0.19           -0.01             0.00
## Albumin                    -0.27           -0.22            -0.23
## Albumin_and_Globulin_Ratio    NA              NA               NA
## Dataset                    -0.14           -0.22            -0.25
##                            Alkaline_Phosphotase Alamine_Aminotransferase
## Age                                        0.08                    -0.09
## Total_Bilirubin                            0.21                     0.21
## Direct_Bilirubin                           0.23                     0.23
## Alkaline_Phosphotase                       1.00                     0.13
## Alamine_Aminotransferase                   0.13                     1.00
## Aspartate_Aminotransferase                 0.17                     0.79
## Total_Protiens                            -0.03                    -0.04
## Albumin                                   -0.17                    -0.03
## Albumin_and_Globulin_Ratio                   NA                       NA
## Dataset                                   -0.18                    -0.16
##                            Aspartate_Aminotransferase Total_Protiens Album
in
## Age                                             -0.02          -0.19   -0.
```

```
27
## Total_Bilirubin                                       0.24            -0.01   -0.
22
## Direct_Bilirubin                                      0.26             0.00   -0.
23
## Alkaline_Phosphotase                                  0.17            -0.03   -0.
17
## Alamine_Aminotransferase                              0.79            -0.04   -0.
03
## Aspartate_Aminotransferase                            1.00            -0.03   -0.
09
## Total_Protiens                                       -0.03             1.00    0.
78
## Albumin                                              -0.09             0.78    1.
00
## Albumin_and_Globulin_Ratio                              NA               NA      
NA
## Dataset                                              -0.15             0.04    0.
16
##                            Albumin_and_Globulin_Ratio Dataset
## Age                                                NA   -0.14
## Total_Bilirubin                                    NA   -0.22
## Direct_Bilirubin                                   NA   -0.25
## Alkaline_Phosphotase                               NA   -0.18
## Alamine_Aminotransferase                           NA   -0.16
## Aspartate_Aminotransferase                         NA   -0.15
## Total_Protiens                                     NA    0.04
## Albumin                                            NA    0.16
## Albumin_and_Globulin_Ratio                          1      NA
## Dataset                                            NA    1.00
```

From the correlation matrix it can be observed that there is a high degree of correlation between the variables ** Total_Bilirubin ** with ** Direct_Bilirubin ** and ** Albumin ** with ** Total_Protiens **, therefore, it can be removed from the database.

```
train_set_mol_2 = train_set %>% select(-Total_Bilirubin,-Total_Protiens)
mol_2 <- glm(Dataset ~. , data = train_set_mol_2, family = binomial())
```

Model 3: Significant variables

```
summary(mol_1)
```

```
## 
## Call:
## glm(formula = Dataset ~ ., family = binomial(), data = train_set)
## 
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1370  -1.0787   0.4107   0.9186   1.4787
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.3869084  1.5202083  -1.570  0.11639
## Age                         0.0193951  0.0069943   2.773  0.00555 **
## GenderMale                 -0.0193057  0.2551699  -0.076  0.93969
## Total_Bilirubin            -0.2239224  0.4940652  -0.453  0.65039
## Direct_Bilirubin            0.9564189  0.9344332   1.024  0.30606
## Alkaline_Phosphotase        0.0006987  0.0008039   0.869  0.38481
## Alamine_Aminotransferase    0.0145580  0.0058760   2.478  0.01323 *
## Aspartate_Aminotransferase  0.0008123  0.0035444   0.229  0.81874
## Total_Protiens              0.5299709  0.4289254   1.236  0.21662
## Albumin                    -0.8757233  0.8457121  -1.035  0.30044
## Albumin_and_Globulin_Ratio  0.6625556  1.2938651   0.512  0.60860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 556.42  on 461  degrees of freedom
## Residual deviance: 459.51  on 451  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 481.51
##
## Number of Fisher Scoring iterations: 7
```

From the summary of model 1, where all the variables were used, it can be seen that the significant ones where their p-value is minus 0.05 are: ** Age ** and ** Alamine_Aminotransferase **. Therefore, only those variables are selected.

```
train_set_mol_3 = train_set %>% select(Age,Alamine_Aminotransferase)
mol_3 <- glm(Dataset ~. , data = train_set, family = binomial())
```

Step 5: Predict the possible ratings for the test base

Model 1: Using all the variables

```
y_hat_1 <- predict(mol_1,test_set,type = "response")
glm.pred_1  <- ifelse(y_hat_1 > 0.5, "1", "0")
accuracy <- data_frame(method="Using all the variables",
                       Accuracy = mean(glm.pred_1 == test_set$Dataset))
```

```r
y_hat_2 <- predict(mol_2,test_set,type = "response")
glm.pred_2  <- ifelse(y_hat_2 > 0.5, "1", "0")
accuracy <- bind_rows(accuracy,data_frame(method="Variables + correlation",
                       Accuracy = mean(glm.pred_2 == test_set$Dataset)))

y_hat_3 <- predict(mol_3,test_set,type = "response")
glm.pred_3  <- ifelse(y_hat_3 > 0.5, "1", "0")
accuracy <- bind_rows(accuracy,data_frame(method="Significant variables",
                                          Accuracy = mean(glm.pred_3 == test_
set$Dataset)))
```

Step 6: Accuracy

```r
accuracy

## # A tibble: 3 x 2
##   method                  Accuracy
##   <chr>                      <dbl>
## 1 Using all the variables    0.769
## 2 Variables + correlation    0.744
## 3 Significant variables      0.769
```

Results

As more relevant results in the exploratory analysis it can be observed that the distribution in men and women for the different variables is similar, this allows us to think that gender is not a relevant variable to take into account, that it can be justified because it does not It is significant in the analysis of the third model. On the basis of the second graph, it can also be observed that certain variables are distributed normally, which is good, if one wishes to make univariate predictions of them. After partitioning the data in training and testing, to be later modeled by means of different approaches, it is evident that through the precision that model 1 and model 3 are equal, take into account that model 3 It only consists of 2 variables (**Age** and **Alamine_Aminotransferase**) to achieve this accuracy, therefore, they are the most important characteristics to know and predict if a patient will suffer from a liver problem.

Conclusion

From the researched literature, decision trees or Vector Support Machine could be taken as models to achieve better prediction levels.